

# Origins of the Combinatorial Basis of Entropy

Robert K. Niven

(1) School of Aerospace, Civil and Mechanical Engineering, The University of New South Wales at ADFA, Canberra, ACT, 2600, Australia. Email: r.niven@adfa.edu.au  
(2) Niels Bohr Institute, University of Copenhagen, Copenhagen Ø, Denmark.

**Abstract.** The combinatorial basis of entropy, given by Boltzmann, can be written  $H = N^{-1} \ln \mathbb{W}$ , where  $H$  is the dimensionless entropy,  $N$  is the number of entities and  $\mathbb{W}$  is number of ways in which a given realization of a system can occur (its statistical weight). This can be broadened to give generalized combinatorial (or probabilistic) definitions of entropy and cross-entropy:  $H = \kappa(\phi(\mathbb{W}) + C)$  and  $D = -\kappa(\phi(\mathbb{P}) + C)$ , where  $\mathbb{P}$  is the probability of a given realization,  $\phi$  is a convenient transformation function,  $\kappa$  is a scaling parameter and  $C$  an arbitrary constant. If  $\mathbb{W}$  or  $\mathbb{P}$  satisfy the multinomial weight or distribution, then using  $\phi(\cdot) = \ln(\cdot)$  and  $\kappa = N^{-1}$ ,  $H$  and  $D$  asymptotically converge to the Shannon and Kullback-Leibler functions. In general, however,  $\mathbb{W}$  or  $\mathbb{P}$  need not be multinomial, nor may they approach an asymptotic limit. In such cases, the entropy or cross-entropy function can be *defined* so that its extremization (“MaxEnt” or “MinXEnt”), subject to the constraints, gives the “most probable” (“MaxProb”) realization of the system. This gives a probabilistic basis for MaxEnt and MinXEnt, independent of any information-theoretic justification.

This work examines the origins of the governing distribution  $\mathbb{P}$ . These include: (a) frequentist-like models; (b) symmetry models; (c) prior MinXEnt models; (d) Kapur-Kesavan inverse models; and (e) game theoretic models. The combinatorial definition and MaxProb are consistent with these different approaches, and the notion of probabilistic inference, yet offer greater utility than traditional MaxEnt / MinXEnt based on the Shannon and Kullback-Leibler functions.

**Keywords:** MaxEnt; MaxProb; Boltzmann principle; combinatorial; probabilistic inference

**PACS:** 02.50.Cw, 02.50.Tt, 05.20.-y, 05.70.-a, 05.90.+m, 89.20.-a, 89.70.+c

## 1. INTRODUCTION

Fifty years ago, Jaynes [1] gave the maximum entropy method (MaxEnt), based on the Shannon entropy [2]:

$$H_{Sh} = - \sum_{i=1}^s p_i \ln p_i \quad (1)$$

where  $p_i$  is the (posterior) probability of occurrence of the  $i$ th distinguishable state within a system, from  $s$  such states. In the MaxEnt method, one maximizes the Shannon entropy of a system, subject to its constraints, to determine the “least informative” or “maximally noncommittal” probability distribution representing the system. From its inception, MaxEnt was advanced as a generic method of inference for the solution of indeterminate problems of all kinds, underpinned by information theory, not merely as an extension of mechanics [1, 3, 4, 5, 6]. MaxEnt was later extended into the maximum relative entropy, minimum divergence or minimum cross-entropy method (MinXEnt), involving extremization of the Kullback-Leibler measure [7, 8]:

$$D_{KL} = \sum_{i=1}^s p_i \ln \frac{p_i}{q_i} \quad (2)$$

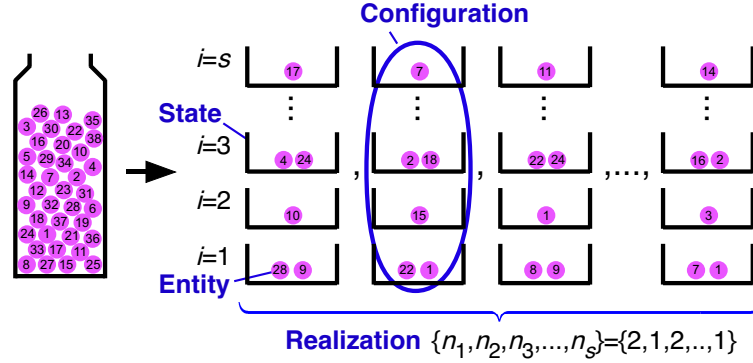
which allows for unequal prior probabilities  $q_i$ . Since that time, MinXEnt and its subsidiary MaxEnt have been successfully applied to the analysis of a vast number of phenomena, throughout most fields of human study [e.g. 6, 9, 10, 11], and can rightly be regarded as one of the most important of all human discoveries.

It must be emphasised, however, that the cross-entropy and entropy concepts which underpin MinXEnt and MaxEnt are themselves subject to many different philosophical interpretations. Dominant explanations include the axiomatic basis outlined by Shannon [2], and the information-theoretic (“bits” of information) and coding basis, recognized by Szilard [12] and Shannon [2] [c.f. 13]. These bases led Jaynes, in particular, to consider the Shannon and Kullback-Leibler functions to be the only logically consistent measures of uncertainty, and thus the only ones suitable for analysis. This view has been challenged by many researchers, on the grounds that the above two measures are too narrowly defined and/or inapplicable to many situations. For example, over the past 85 years, many alternative entropy and divergence functions have been introduced [e.g. 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]; in most cases, these are incompatible with the Shannon and Kullback-Leibler functions, but have proved *useful* for the analysis of specific classes of systems. Can such measures be explained by some broader philosophical framework? How should we choose the “correct” cross-entropy or entropy function for a given problem? The fact that such questions remain unanswered indicates the need for a unifying philosophical framework, which encompasses (and *explains*) such alternative entropy measures and their connections to information theory.

This study examines one such framework: the combinatorial (or probabilistic) basis of entropy, first given 130 years ago by Boltzmann [31] and subsequently promoted by Planck [32]. This involves the maximization of a governing probability distribution  $\mathbb{P}$  or weight  $\mathbb{W}$  of a system; this can be viewed as a generalized principle of probabilistic inference, aptly described by Vincze and Grendar & Grendar as the maximum probability (“MaxProb”) principle [33, 34]. It also leads to generalized definitions of cross-entropy and entropy, based purely on probability theory [35]. In this study, specific attention is paid to the origins of the governing distribution  $\mathbb{P}$ , including (a) frequentist-like models (e.g. ball-in-box or urn models); (b) symmetry models; (c) prior MinXEnt models; (d) Kapur-Kesavan inverse models; and (e) game theoretic models. It is shown that the combinatorial basis is consistent with these different approaches, but is more soundly based and offers greater utility than traditional MaxEnt / MinXEnt based on the Shannon and Kullback-Leibler functions.

## 2. THE COMBINATORIAL BASIS

Owing to a tremendous confusion in terminology - especially amongst physicists - it is first necessary to rigorously define several important terms [c.f. 35]. An *entity* is here taken to be a discrete particle, object or agent within a system, which acts separately but not necessarily independently of the other entities present. A *system* is a collection of entities with a defined boundary, subject to various constraints, which may or may not



**FIGURE 1.** Definition of terms used in the combinatorial basis of entropy and cross-entropy.

be open to the exchange of specified entities or substances with an external environment. The entity therefore constitutes the unit of analysis of a system.

Now consider a simple “ball-in-box” model of a system, shown in Figure 1, in which  $N$  distinguishable entities (balls) are allocated to  $s$  distinguishable non-degenerate states (boxes). As shown:

- A *state* refers to each different category or element of system (e.g. energy levels, sides of a die or alphabetic symbols). The states are therefore properties of, or associated with, each individual entity in the system.
- A *configuration* is a distinguishable permutation or pattern of entities amongst the states of a system (a *complexion*, *microstate* or *sequence*). A configuration is therefore a property of the system as a whole.
- A *realization* is each aggregated arrangement of entities amongst the states of a system, as specified by some rule, for example by the number of entities in each state (a *macro-state*, *outcome* or *type*). In general, a realization will constitute a set of configurations, since several configurations could give the same realization (see Figure 1).

There is such confusion in and sloppy usage of the terms *state*, *microstate* and *macro-state* - severely impairing understanding - that the last two terms should be avoided. In the following, the states are indexed  $i = 1, \dots, s$  (which may be multivariate);  $n_i$  denotes the number of entities in the  $i$ th state;  $q_i$  and  $p_i = n_i/N$  respectively denote the prior and posterior probabilities of an entity being in the  $i$ th state; and each realization<sup>1</sup> is denoted  $\{n_i\}$ . Notwithstanding other philosophical differences with Jaynes, the “subjective Bayesian” definition of probabilities, as assignments based on what we know, is adopted here [1, 6].

For the analysis of probabilistic systems, it is possible to delineate a principle which stands out from all others: the *maximum probability* (“MaxProb”) principle [31, 32, 33, 34, 35]. This can be stated as:

“A system can be represented by its realization of highest probability.”

<sup>1</sup> A realization can only be denoted  $\{p_i\}$  in the asymptotic limits  $N \rightarrow \infty$  and  $n_i \rightarrow \infty, \forall i$ , since  $\{p_i\}$  discards information about the value of  $N$ .

This seemingly trivial statement provides a powerful principle for *probabilistic inference*, which is independent of any information-theoretic considerations. This is critical, since in any contradiction between information theory and probability theory - for example, between the distributions inferred by each approach - probability theory must triumph. Like MinXEnt or MaxEnt based on the Kullback-Leibler or Shannon measures, MaxProb is a method of inference (inductive reasoning), which does not give certainty in its predictions. Unlike them, however, MaxProb is founded solely on probability theory. Indeed, MaxProb does not depend upon any asymptotic limits (a feature of the “frequentist” definition of probability, in which probabilities must correspond to measurable frequencies [1, 6]); it can therefore be applied to systems containing finite numbers of entities [29, 30].

Allied to MaxProb is a generalized form of the second law of thermodynamics:

*“A system will tend towards its most probable realization.”*

This provides a purely probabilistic rationale for use of the MaxProb principle, independent of thermodynamics. In effect, if we adopt MaxProb as a principle for probabilistic inference, the above statement is its corresponding ergodic principle, which (on average) explains its success. Of course - as expressed by Jaynes [1] - the concept of ergodicity is not needed for the purpose of inference, since in the absence of other information, we are fully justified in conducting inference without it.

The MaxProb principle also leads to the *combinatorial definition* of entropy, first given by Boltzmann [31] and Planck [32]. This can be written as:

$$H = N^{-1} \ln \mathbb{W}, \quad (3)$$

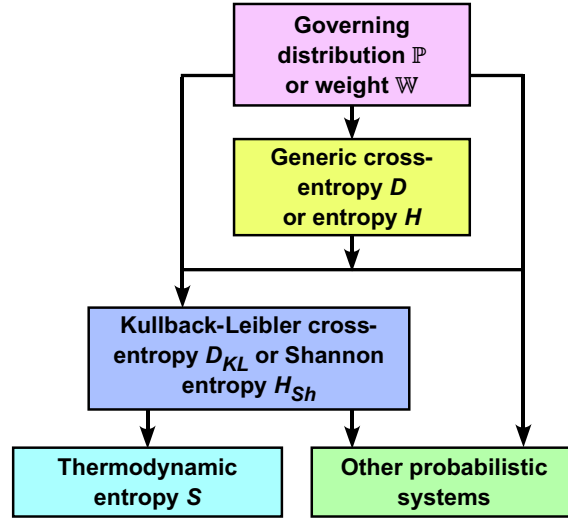
where  $\mathbb{W}$  is the number of ways in which a given realization can occur, referred to as its statistical weight. Maximization of the entropy  $H$  of a system, subject to its constraints, therefore selects the realization of highest weight  $\mathbb{W}$  (the logarithmic function being a monotonic transformation, which does not alter the position of the extremum). Eq. (3) can be extended to give generalized combinatorial (or probabilistic) definitions of cross-entropy and entropy [35]:

$$D = -\kappa(\phi(\mathbb{P}) + C), \quad H = \kappa(\phi(\mathbb{W}) + C), \quad (4)$$

where  $\mathbb{P} = P(\{n_i\}|\{q_i\}, N, s, I)$  is the probability of a given realization, subject to the prior probabilities  $\{q_i\}$ , number of entities  $N$ , number of states  $s$  and background information  $I$ ;  $\phi$  is a convenient monotonic transformation function;  $\kappa$  is a scaling parameter; and  $C$  is an arbitrary constant. This perspective is summarised in Figure 2. If  $\mathbb{P}$  or  $\mathbb{W}$  satisfy the multinomial distribution or weight:

$$\mathbb{P} = N! \prod_{i=1}^s \frac{q_i^{n_i}}{n_i!}, \quad \mathbb{W} = N! \prod_{i=1}^s \frac{1}{n_i!}, \quad (5)$$

then by taking  $\phi(\cdot) = \ln(\cdot)$ ,  $\kappa = N^{-1}$  and the asymptotic limits  $N \rightarrow \infty$  and  $n_i \rightarrow \infty, \forall i$  (the “Stirling approximation”),  $D$  and  $H$  converge respectively to the Kullback-Leibler and Shannon functions (1)-(2) [33, 34]. This provides a (well-known) justification for these functions, and their corresponding MinXEnt and MaxEnt principles, as a special case, independently of the arguments used in information theory.



**FIGURE 2.** Schematic flowchart of the combinatorial basis of entropy and cross-entropy.

In general, however,  $\mathbb{P}$  or  $\mathbb{W}$  need not be multinomial, nor may they approach an asymptotic limiting form. In such cases, extremization of the cross-entropy or entropy defined by (4), subject to the constraints, gives the most probable (MaxProb) realization of the system (in the non-asymptotic case, due to the effect of quantization, extremization gives an “attractor” distribution which lies close to but not necessarily equal to the MaxProb realization [35]). In consequence, the combinatorial definitions (4) remain consistent with the rules of probability theory, whilst inference using the Kullback-Leibler or Shannon measures may lead to inconsistencies. The combinatorial (or probabilistic) definitions are therefore more broadly applicable than those derived from information-theoretic considerations.

In the foregoing discussion, the astute reader will notice that there may be many different ways to classify the entities and states of a system, and hence to identify its configurations; and many different ways to group the configurations into realizations. We are therefore led to the “subjective” (or “observer-dependent”) view of the entropy and cross-entropy concepts, a sentiment vocally defended by Jaynes [1]. This was succinctly expressed by Tseng and Caticha [36]:

*“Entropy is not a property of a system ... [it] is a property of our description of a system.”*

The fact that the thermodynamic entropy  $S$  is always defined in the same manner, allowing thermodynamicists to make consistent calculations, should not fool the reader into believing that the entropy concept is “objective”<sup>2</sup>.

<sup>2</sup> It is important that the symbol  $S$  be devoted exclusively to the thermodynamic entropy, since it is a special case of - but is distinct from - the dimensionless Shannon entropy (1).

### 3. ORIGINS OF THE GOVERNING DISTRIBUTION

We now consider the origins of the governing distribution  $\mathbb{P}$  or weight  $\mathbb{W}$  used in the combinatorial formulation. The problem of justifying a cross-entropy or entropy function is now replaced by a deeper problem, of how to justify its governing distribution. The ubiquity of the Kullback-Leibler and Shannon measures, in many circumstances, therefore leads to the question: why the multinomial distribution? This question, and the choice of  $\mathbb{P}$ , is examined from five different perspectives.

#### (a) Frequentist-like Models

In this approach, one simply asserts a governing distribution  $\mathbb{P}$  or weight  $\mathbb{W}$  as a probabilistic model of the system under consideration. One may have strong grounds, based on prior knowledge of a problem, for such an assertion; in any case, we should have no “fear of failure” of this method (in Jaynes’ words), since if the model gives unsuccessful predictions, we have learnt that it is incorrect. Many such models are available from classical probability theory, for example the *ball-in-box* models of the type represented in Figure 1. Since these arose from frequentist studies, they can be termed “frequentist-like” models, although used here for the purpose of inference.

In the case discussed previously, in which distinguishable balls are allocated to distinguishable boxes in accordance with a set of constant prior probabilities (see Figure 1), one obtains the multinomial distribution (5), and hence the Kullback-Leibler cross-entropy and Shannon entropy functions in the Stirling limits. However, different assumptions lead to different model distributions. If the asymptotic limits are not applied, then from (5), one obtains a non-asymptotic cross-entropy function [c.f. 29]:

$$\begin{aligned} -D_{KL}^x &= N^{-1} \ln \mathbb{P} = N^{-1} \left\{ \ln N! + \sum_{i=1}^s n_i \ln q_i - \sum_{i=1}^s \ln n_i! \right\} \\ &= \sum_{i=1}^s \left\{ p_i N^{-1} \ln N! + p_i \ln q_i - N^{-1} \ln[(p_i N)!] \right\} \end{aligned} \quad (6)$$

This is applicable to systems with finite (small)  $N$ . Minimisation of  $D_{KL}^x$ , subject to the usual constraints  $\sum_{i=1}^s n_i = N$  and  $\sum_{i=1}^s n_i f_{ri} = N \langle f_r \rangle$ , for  $r = 1, \dots, R$ , where  $f_{ri}$  is the  $r$ th function of each state  $i$  and  $\langle f_r \rangle$  is its mathematical expectation, gives the “most probable” distribution [c.f. 29]:

$$p_i^\# = N^{-1} \left[ \psi^{-1} \left( N^{-1} \ln N! + \ln q_i - \lambda_0 - \sum_{r=1}^R \lambda_r f_{ri} \right) - 1 \right] \quad (7)$$

where  $\psi^{-1}(\cdot)$  is the inverse digamma function. Eq. (7) can be viewed as the “attractor” for systems with finite  $N$ , which differs from the attractor given by traditional MinXEnt.

If the states are considered to contain  $g_i$  distinguishable, degenerate sub-states within each distinguishable state  $i$ , then three cases have been examined historically: (i) distinguishable entities; (ii) indistinguishable entities; and (iii) indistinguishable entities, with a maximum of one entity in each state. The resulting distributions were given by Brillouin [37, 38] as, respectively:

$$\mathbb{P}_{MB} = \frac{N!}{G^N} \prod_{i=1}^s \frac{g_i^{n_i}}{n_i!}, \quad (8)$$

$$\mathbb{P}_{BE} = \frac{N!(G-1)!}{(G+N-1)!} \prod_{i=1}^s \frac{(g_i+n_i-1)!}{n_i!(g_i-1)!}, \quad (9)$$

$$\mathbb{P}_{FD} = \frac{N!(G-N)!}{G!} \prod_{i=1}^s \frac{g_i!}{n_i!(g_i-n_i)!}. \quad (10)$$

where  $G = \sum_{i=1}^s g_i$  is the total degeneracy. The truncated weights and entropy functions corresponding to these distributions, referred to respectively as the Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac distributions respectively [e.g 16, 17, 18, 19, 20, 37, 38, 39, 40, 41], played an important role in the development of quantum theory. In the non-asymptotic case, the resulting entropy functions appear to have profound information-theoretic consequences [29, 30].

Recently, a quite different ball-in-box model was considered, in which distinguishable entities are allocated to indistinguishable, equally degenerate states. The statistical weight of each realization  $\{n_i\}$  can be expressed as [42]:

$$\mathbb{W}_{D:I(g)} = \frac{N!}{\left(\prod_{i=1}^k n_i!\right) \left(\prod_{j=1}^N r_j!\right)} \prod_{i=1}^k \sum_{\gamma=1}^{\min(g, n_i)} \left\{ \begin{matrix} n_i \\ \gamma \end{matrix} \right\} \quad (11)$$

where there are  $k$  non-empty states amongst the  $s$  states;  $g$  is the degeneracy of each state;  $\left\{ \begin{matrix} n_i \\ \gamma \end{matrix} \right\}$  is a Stirling number of the second kind; and  $r_j$  is the number of occurrences of integer  $j$  in the set  $\{n_i\}$ . The combinatorial entropy corresponding to (11),  $H_{D:I(g)} = N^{-1} \ln \mathbb{W}_{D:I(g)}$ , does not appear to have a straightforward asymptotic form, except in the non-degenerate case  $g = 1$  with  $k = s$ , when it reduces to the Shannon entropy.

Closely related to but distinct from ball-in-box models are *urn models*, in which a container (urn) is set up with a total of  $M$  balls, made up of  $m_i$  balls of each color  $i$ . Balls are then drawn from the urn in accordance with some sampling scheme, recorded and returned to the urn (or the urn modified in some way), and the sampling repeated [c.f. 43, 44]. The asymptotic limits of an infinitely large urn ( $M \rightarrow \infty$  and  $m_i \rightarrow \infty, \forall i$ ), and an infinitely large (smaller) sample ( $N \rightarrow \infty$  and  $n_i \rightarrow \infty, \forall i$ ), are usually applied. Although quite different to the ball-in-box model of Figure 1, an urn model with simple replacement also yields the multinomial distribution [43, 44]. Urn models involving the drawing of balls without replacement, or double replacement, lead respectively to the Fermi-Dirac and Bose-Einstein distributions [43]. Urn models also readily permit the construction of systems in which the prior probabilities are not independently and identically distributed (non-*iid* sampling): e.g. the Pólya distribution, in which after every draw, the ball is returned, and  $c$  balls of the same color are also added [45, 46, 47, 48]:

$$\mathbb{P}_{Polya} = \frac{N!}{\prod_{i=1}^s n_i!} \prod_{i=1}^s \frac{m_i(m_i+c) \dots (m_i+(n_i-1)c)}{M(M+c) \dots (M+(N-1)c)}, \quad (12)$$

Substituting the *initial* prior probabilities  $q_i = m_i/M$  and parameter  $\beta = N/M$ , this gives analytic cross-entropy measures in the non-asymptotic and asymptotic cases [48]. The resulting “most probable” distribution is intermediate between the Bose-Einstein and Fermi-Dirac distributions, with physical applications.

### **(b) Symmetry-Based Arguments**

One may also choose a governing distribution on the basis of symmetry arguments (related to the “principle of insufficient reason”). For a system made up of tosses of a coin, it is rational to consider the sampling to follow the binomial distribution, with equal prior probabilities of  $\frac{1}{2}$  for each face, due to the symmetry of the states (there being no information to suggest that one state should be preferred). Alternatively, as suggested by David Blower at MaxEnt07, one can obtain a binomial distribution by the symmetry of all possible models in the model space (assigning a uniform prior to the models, over the entire spectrum from an all-head to an all-tail model, there being no reason to prefer any model). Applied to systems with more than two states, either argument leads to the multinomial distribution. In this respect, the multinomial distribution plays a role somewhat analogous to a central limit theorem (a “central model theorem”), a point which deserves greater mathematical attention; this may be the reason for the ubiquity of the Kullback-Leibler and Shannon measures. Without symmetry, however, the argument breaks down, and one must adopt some other method to identify the governing distribution.

### **(c) Prior MinXEnt Models**

A third origin of the governing distribution  $\mathbb{P}$  is as a result of the application of MinXEnt at a higher level, for example to the set of systems within which the actual system resides. For example, the multinomial distribution can be obtained by MinXEnt based on the Kullback-Leibler cross-entropy, subject to a multinomial prior and mean constraints on each variate [11]. This might then be imposed as a lower-level governing distribution. One can in fact envisage a hierarchy of governing and “most probable” distributions, at different levels of description. In a complex system, in which there is bidirectional feedback, the result will be a mosaic of interconnected probabilistic models (with thanks to the discussion by Tony Bell at MaxEnt07).

### **(d) Kapur-Kesavan Inverse Models**

The governing distribution  $\mathbb{P}$  can also be obtained by extension of the arguments of Kapur and Kesavan [11], in which one works backwards from an observed probability distribution  $\{p_i^*\}$ , prior distribution  $\{q_i\}$  (if available) and any constraints, to derive the measure of cross-entropy or entropy applicable to a system. By unravelling of the asymptotic limits, this could (at least in principle) be extended to determine the governing distribution of the system. This avenue of research has not been examined in detail, and deserves greater attention.

### **(e) Game Theoretic Models**

The final method considered here is to derive the governing distribution of a system by analysis of a code-length game between the system (“Nature”) and the observer [49, 50]. For a multivariate system of *iid* random variables, which take discrete values, this yields the multinomial distribution at game-theoretic equilibrium [49]. As in case (c), this could then be imposed as the governing distribution at a lower level of description.



## 4. CONCLUSIONS

This study examines the MaxProb principle, in which a system is represented by its distribution of highest probability. This can be interpreted as a generalized method of probabilistic inference, which does not provide certainty in its predictions, yet is always consistent with the rules of probability theory. In contrast, inference using the Kullback-Leibler cross-entropy or Shannon entropy functions, in cases in which the governing distribution is not multinomial and/or does not satisfy the asymptotic limits, can lead to inconsistencies. The MaxProb principle also gives rise to generalized combinatorial definitions of cross-entropy and entropy, an extension of the idea given by Boltzmann 130 years ago. The cross-entropy or entropy can therefore be *defined* so that its extremization, subject to the constraints, gives the “most probable” (“MaxProb”) realization of the system. This provides a purely probabilistic basis for MaxEnt and MinXEnt, which is independent of any information-theoretic justification.

This work examines the origins of the governing distribution  $\mathbb{P}$ , including by (a) frequentist-like models (e.g. ball-in-box or urn models); (b) symmetry models; (c) prior MinXEnt models; (d) Kapur-Kesavan inverse models; and (e) game theoretic models. It is shown that the combinatorial definition and MaxProb are consistent with these different approaches, and the “subjective Bayesian” definition of probability, yet is more broadly based and offers greater utility than traditional MaxEnt / MinXEnt based on the Shannon and Kullback-Leibler functions.

## ACKNOWLEDGMENTS

The author thanks the European Commission for support as a Marie Curie Incoming International Fellow; Marian Grendár, David Blower, Tony Bell and Flemming Topsøe for specific arguments (as detailed above) and stimulating discussions; and the organisers and participants of MaxEnt07.

## REFERENCES

1. E.T. Jaynes, Phys. Rev. 106 (1957) 620.
2. C.E. Shannon, Bell Sys. Tech. J. 27 (1948) 379; 623.
3. E.T. Jaynes, in K.W. Ford (ed.), Brandeis University Summer Institute, Lectures in Theoretical Physics, Vol. 3, Benjamin-Cummings Publ. Co. (1963) 181.
4. E.T. Jaynes, IEEE Trans. Systems Science and Cybernetics SSC-4 (1968) 227.
5. E.T. Jaynes, in R.D. Levine, M. Tribus (eds.), The Maximum Entropy Formalism, MIT Press, Cambridge, MA (1978) 15.
6. E.T. Jaynes, (G.L. Bretthorst, ed.) Probability Theory: The Logic of Science, Cambridge U.P., Cambridge, 2003.
7. S. Kullback, R.A. Leibler, Annals Math. Stat. 22 (1951) 79.
8. S. Kullback, Information Theory and Statistics, John Wiley, NY, 1959.
9. R.D. Levine, M. Tribus (eds.), The Maximum Entropy Formalism, MIT Press, Cambridge, MA, 1978.
10. J.N. Kapur, Maximum-Entropy Models in Science and Engineering, John Wiley, NY, 1989.
11. J.N. Kapur, H.K. Kesavan, Entropy Optimization Principles with Applications, Academic Press, Inc., Boston, MA, 1992.

12. L. Szilard, Zeitschrift für Physik 53 (1929) 840.
13. J.E. Shore, R.W. Johnson, IEEE Trans. Information Theory IT-26(1) (1980) 26.
14. R.A. Fisher, Philos. Trans. Royal Soc. London A 222 (1922) 309.
15. R.A. Fisher, Proc. Camb. Philos. Soc. 22 (1925) 700.
16. S.N. Bose, Z. Phys. 26 (1924) 178.
17. A. Einstein, Sitzungsber. Preuss. Akad. Wiss. Phys. Math. Kl (1924) 261.
18. A. Einstein, Sitzungsber. Preuss. Akad. Wiss. Phys. Math. Kl (1925) 3.
19. E. Fermi, Z. Phys. 36 (1926) 902.
20. P.A.M. Dirac, Proc. Roy. Soc. 112 (1926) 661.
21. A. Rényi, Proc. 4th Berkeley Symp. Math. Stat. and Prob. 1 (1961) 547.
22. B.D. Sharma, D.P. Mittal, J. Math. Sci. (Calcutta) 10 (1975) 28.
23. B.D. Sharma, D.P. Mittal, J. Combinat. Inform. Sys. Sci. 2 (1977) 122.
24. C. Tsallis, J. Stat. Phys. 52(1/2) (1988) 479.
25. C. Tsallis, in S. Abe, Y. Okamoto (eds.), Nonextensive Statistical Mechanics and its Applications, Springer, Berlin, (2001) 3.
26. G. Kaniadakis, Physica A 296(3-4) (2001) 405.
27. G. Kaniadakis, Phys. Rev. E 66(5) (2002) 056125.
28. C. Beck, E.G.D. Cohen, Physica A 322 (2003) 267.
29. R.K. Niven, Phys. Lett. A 342(4) (2005) 286.
30. R.K. Niven, Physica A 365(1) (2006) 142.
31. L. Boltzmann, Wiener Berichte, 76 (1877) 373-435.
32. M. Planck, Annalen der Physik 4 (1901) 553.
33. I. Vincze, Progress in Statistics, 2 (1974) 869-895.
34. M. Grendár, Jr. and M. Grendár, in Bayesian Inference and Maximum Entropy Methods in Science and Engineering, A. Mohammad-Djafari (ed.), AIP, Melville (2001) 83-94.
35. R. K. Niven, *cond-mat/0512017*, 2005-2007.
36. C.-Y. Tseng, A. Caticha, Yet another resolution of the Gibbs paradox: an information theory approach, preprint (2002).
37. L. Brillouin, Annales de Physique 7 (1927) 315.
38. L. Brillouin, Les Statistiques Quantiques et Leurs Applications, Les Presses Universitaires de France, Paris, 1930.
39. R.C. Tolman, The Principles of Statistical Mechanics, Oxford Univ. Press, London, 1938.
40. L. Brillouin, J. Appl. Phys. 22(3) (1951) 338.
41. N. Davidson, Statistical Mechanics, McGraw-Hill, NY, 1962.
42. R.K. Niven, CTNEXT07, Catania, Sicily, Italy, July 2007, in submission to AIP.
43. D.R. Jensen, in S. Kotz, N.L. Johnson, Encyclopedia of Statistical Sciences, 6: 5200.
44. S. Berg, Urn Models, in S. Kotz, N.L. Johnson, *Encyclopedia of Statistical Sciences*, 9: 424.
45. F. Eggenberger, G. Pólya, Über die Statistik verketteter Vorgänge Z. Angew. Math. Mech., 1 (1923) 279-289.
46. H. S. Steyn, Proc. Koninklijke Nderlandse Akademie van Wetenschappen, Ser. A, 54 (1951) 23-30.
47. N. L. Johnson, S. Kotz, N. Balakrishnan, Discrete Multivariate Distributions. New York: Wiley, 1997.
48. M. Grendar, R.K. Niven, *cond-mat/0612697*, 2006.
49. F. Topsøe, IEEE Trans. Info. Theory 48(8) (2002) 2368.
50. F. Topsøe, Physica A 340 (2004) 11.